图书情报工作
第 64 卷 第 18 期　2020 年 9 月

一种基于图卷积自编码模型的多维度学科知识网络融合方法*

■ 李慧　胡吉霞

西安电子科技大学经济与管理学院　西安 710126

摘　要：[目的/意义]针对包含单一类型知识单元的知识网络难以全面反映学科知识结构的问题，提出一种从多维度进行知识网络结构融合的方法，为学科领域知识结构挖掘提供借鉴。[方法/过程]利用 LDA 及 TF-IDF 方法抽取学科知识单元，然后运用语义相似度和关键词共现分析方法构建 3 个学科知识子网络：主题网络、关键词网络和实体网络，并采用空间节点传递对齐方法对齐子网络节点，接着设计基于图卷积操作的自编码模型对知识节点进行表示，最后通过计算余弦相似度重构学科知识网络。[结果/结论]实验部分以人工智能领域为例，构建融合主题、关键词和实体的学科知识网络并展开分析，实验结果表明，本文所提方法能有效地揭示学科领域研究内容和知识结构，为学科知识发现与组织研究提供有益参考。

关键词：网络融合　知识结构　节点对齐　图卷积神经网络　自编码模型

分类号：G254

DOI：10.13266/j.issn.0252-3116.2020.18.013

1 引言

互联网技术的快速发展使得信息的增长速度十分迅猛，在为人们搜集和获取知识信息带来便利的同时，也使人们陷入大量分散、多样化的知识信息海洋中，给从宏观上全面了解知识信息结构带来一定的困扰。随着科学研究范围的不断扩大和研究内容的不断深入，各领域的知识信息呈现出交叉融合的复杂局面。在这种情况下，一个刚进入某领域的学者想要快速了解领域知识结构和发展现状存在一定的困难，如何梳理学科知识结构、有效组织领域知识信息成为亟待解决的问题。学科知识结构是利用不同知识组织方式，揭示知识的本质及彼此间的关联，根据不同的关联类型[1]和表现形式可概括为层次知识结构和网状知识结构，与层次知识结构相比，网状的学科知识网络[2]以其丰富直观的知识表示方式受到学者的广泛关注。构建学科知识网络为领域知识信息组织和知识结构呈现提供了有效途径，分析学科知识网络已成为挖掘学科知识结构和探测学科领域前沿的重要方法和手段。学科知识网络不仅可以从微观层面揭示科学知识网络中知识节点间的相互关系，还可以反映科学概念和热点研究在领域中的变化规律[3]，跟踪新兴专业学科领域知识结构变化态势对科研管理者、科学研究者和政策制定者的工作具有十分重要的意义。

近年来，学科知识网络的研究成果不断涌现，通过对相关文献的梳理发现，目前关于学科知识网络的研究，多在文献计量学的基础上，以作者、机构、期刊、引文、主题、关键词等单一知识单元构建学科知识网络，侧重于学科热点发现及领域合作情况的研究，无法完整地揭示某一学科研究的内在知识结构。针对传统方法的不足，本文以领域科技文献为研究对象，提出一种整合主题、关键词和实体的学科知识网络构建方法，该方法首先抽取主题、关键词和实体作为知识单元，基于关键词共现分析和语义相似度计算方法，构建各维度知识关联子网络，然后利用节点聚类和图卷积自编码模型挖掘知识单元之间更深层次的语义信息和结构信

* 本文系中央高校基本科研业务费专项资金项目"专利视角下的技术创新主题发现与趋势预测"（项目编号：JB190610）研究成果之一。

作者简介：李慧（ORCID：0000-0002-3468-5170），副教授，博士，硕士生导师；胡吉霞（ORCID：0000-0003-3864-2562），硕士研究生，通讯作者，E-mail：2465541453@qq.com。

收稿日期：2020-02-21　修回日期：2020-04-13　本文起止页码：114-125　本文责任编辑：王传清

李慧，胡吉霞. 一种基于图卷积自编码模型的多维度学科知识网络融合方法[J]. 图书情报工作,2020,64(18):114－125.

息,生成一个融合主题、关键词和实体的学科知识网络。以多维度学科研究的内容信息来构建学科知识网络,突破了传统方法以单一维度知识单元的共现关系刻画知识结构的局限性,能够更完整地体现领域知识单元之间的关联情况,从而达到全面准确地揭示学科知识结构的目的。

2 相关研究

近年来,关于学科知识结构理论和方法的研究成果不断涌现,具体归纳如下。

2.1 理论研究

关于知识网络的研究层出不穷,知识网络的基础理论也在不断丰富,但对知识网络的定义没有统一界定。A. Seufert 等认为知识网络是由行为主体、主体之间的关系以及所运用的资源和制度特性 3 个方面组成的动态框架,通过知识转移和知识创造过程积累和运用知识,最终实现价值创造[4]。赵蓉英将知识网络抽象概括为:以知识元素、知识点、知识单元、知识库作为“节点”,以知识间的关联作为“边”或“链”而构成的网络[5]。顾东蕾以哲学的方法描述学科知识网络的内涵,认为学科知识网络是由学科知识元素组成的知识节点和知识关联(知识链接)构成的网状知识体系,具有知识场分布性、相对真伪性和有序性[6]。另外,不同领域对知识网络的内涵和外延有不同的认识。从图书情报领域知识组织的角度来看,知识网络是由学科领域知识节点和知识关联构成的网络[7]。国外的管理学界将知识网络定义为“是一批人、资源和他们之间的关系,为了知识的积累和利用,通过知识创造、知识转移,促进新的知识的利用”[8]。社会学领域则认为知识网络是一种“人际关系网络”,人们可以从中获取或交换物质、信息、知识或情报等资源。虽然知识网络没有统一的定义,但各种知识网络定义的内涵都可以理解为知识网络的主体(节点)之间的交互(关系)作用。

2.2 方法研究

近年来,随着复杂网络分析方法和技术的兴起,目前关于知识网络构建的方法研究,大多数是以文献计量学和社会网络分析(Social Network Analysis, SNA)为基础进行展开。利用社会网络分析方法构建学科知识网络是在文献计量的基础上,以文献外部特征如作者、机构、期刊、引文等的共现关系生成特定领域科学合作网络[9－10]、共被引网络[11－13]等,通过对知识网络进行属性分析、中心性分析、核心－边缘结构分析、凝聚子群分析、节点聚类、关键节点识别等方法发现领域研究热点和探测领域前沿,这类基于共现的方法分析学科知识结构的研究成果十分丰富。随着研究范围的不断扩大,有学者深入文献内部语义特征如文献标题、摘要、关键词、全文等,深度挖掘学科内部知识结构,并且在农业[14]、经济[15]、医学[16]等领域得到广泛应用。吕鹏辉等[17－19]分别总结了引文网络、共被引网络和共词网络的结构、特征和演化研究方法、程序和图谱绘制流程,揭示了知识网络节点之间的关系,并对相应知识网络研究方法的局限性进行了讨论。关鹏、王曰芬等[20－22]提出了整合主题的学科知识网络构建与分析框架,扩展了学科知识网络的研究范围,又利用主题在科学文献中的共现关系构建主题－主题关联的学科知识网络,提出了主题影响力概念和度量方法,后来又基于作者－主题模型构建作者－主题关联的二模学科知识网络,利用作者在网络中的中心性指标度量作者主题关联影响力,弥补了单个引文网络和作者合著网络分析的不足。

综上所述,当前关于学科知识网络的研究主要涉及概念等理论方面的探讨以及基于文献外部属性和单一内容信息构建学科知识网络,在内容层面整合多维知识单元构建知识网络的研究较少,并且鲜有将实体作为学科知识单元的研究成果。因此,本文采用图卷积神经网络方法,尝试设计整合主题、关键词和实体的学科知识网络构建框架,扩展学科知识网络构建方法,也为揭示学科知识结构提供新途径。

3 多维度知识网络融合方法研究

3.1 研究框架

本文综合运用语义相似度和图卷积神经网络计算方法,设计主题、关键词、实体这 3 个学科知识子网络的生成方法和整合模型,以期全面准确地揭示学科领域知识结构,具体流程如图 1 所示。首先,对原始数据集进行预处理生成语料库,抽取学科知识单元,包括主题、关键词和实体,基于知识单元之间的语义相似度构建各维度知识关联子网络;其次,对所有节点进行聚类生成模板网络,用节点传递对齐方法[23]将各子网络转化为节点固定大小的网络结构;最后,利用图卷积运算结合自编码模型融合各子网络并可视化,清晰明了地揭示学科知识分布及关联情况。

3.2 知识单元抽取

科技文献作为科学研究活动最直接有效的表现形式,其内容蕴含和承载了不同学科领域的研究主题、动态演化脉络和发展趋势[24],对科学研究具有重要的参考

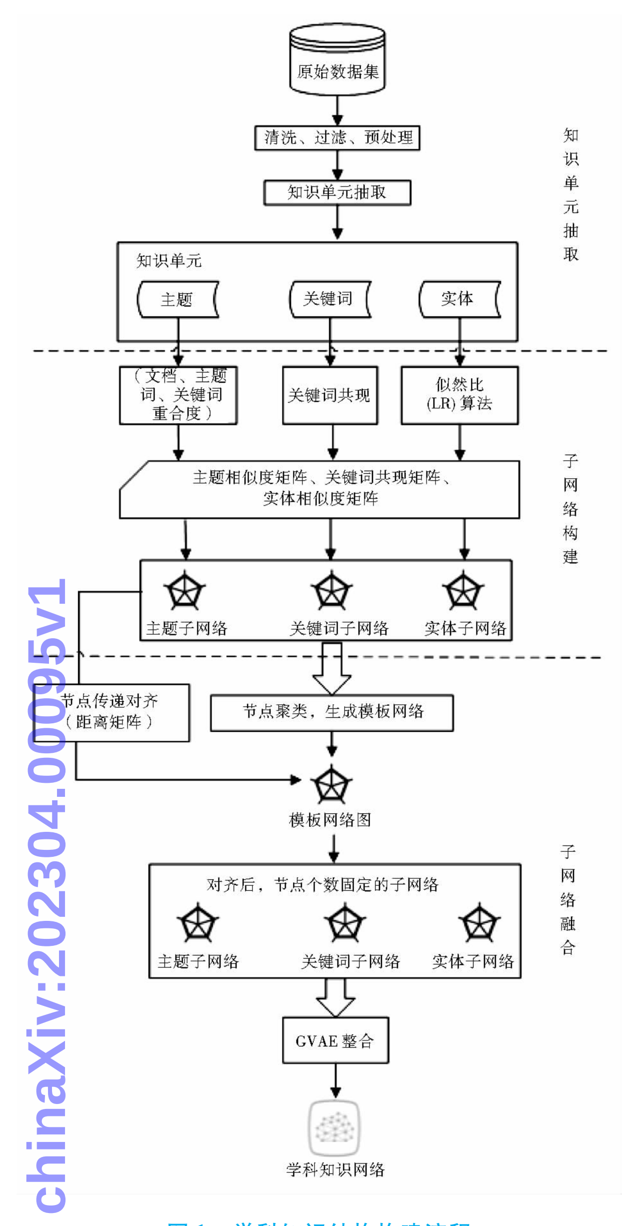


图 1 学科知识结构构建流程

价值和指导作用。本文以科技文献作为初始数据来源，从中抽取能反映学科知识的主题、关键词和实体，建立学科知识单元表征体系。各维度知识单元抽取方法如下：

3.2.1 主题

本文利用 LDA[25] 主题模型进行主题抽取，首先对原始数据集进行清洗、分词等预处理，然后用困惑度确定最优主题数，设定 LDA 模型参数并运行 LDA 程序，最后对程序输出的“主题－词”文件进行概括总结，由此得到学科知识主题。

3.2.2 关键词

关键词是对文献内容的高度凝练和概括，能在更大程度上反映学科研究内容，关键词可直接从语料库中提取。

3.2.3 实体

利用改进的 TF-IDF[26] 算法（见公式（1）和公式（2））筛选出语料库中的高频词，进行词性标注后，选择 TF-IDF 值较大的 N 个名词作为实体，参考主题和关键词的个数确定 N 的取值。

$$\text{TF-IDF}_{wd} = \frac{TF_{wd}^{2} + IDF_{w}}{|d|} \qquad \text{公式(1)}$$

$$\text{IDF}_{w} = \log\left(\frac{N}{DF_{w}+1}\right) \qquad \text{公式(2)}$$

其中，w 表示词语，d 表示文档，wd 表示词语 w 在文档 d 中的频次，|d| 表示文档 d 中包含词语的数量。

3.3 知识子网络构建

以 3.2 节抽取的主题、关键词和实体作为知识节点，以知识节点之间的相似程度确定连边，分别构建主题子网络、关键词子网络和实体子网络。综合文档重合度[27]、主题词相似度和关键词相似度计算主题相似度，再分别利用关键词共现分析和 LR（Likelihood Ratio）[26] 方法计算关键词相似度和实体相似度，确定合适的阈值，选择相似度大于阈值的节点连边构建相似度矩阵来挖掘知识单元之间的关系特征，以此构建学科知识子网络。

3.3.1 主题子网络

将文档重合度（$doc_{coincidence}$）、主题词相似度（$feature_{sim}$）和关键词相似度（$keywords_{sim}$）3 个指标的加权和作为主题之间相似程度的度量指标，如公式（3）所示。各指标计算方法如下：

$$topic_{sim}(i,j) = w_1\, doc_{coincidence(i,j)} + w_2\, feature_{sim(i,j)} + w_3\, keywords_{sim(i,j)} \qquad \text{公式(3)}$$

其中，$topic_{sim(i,j)}$ 表示主题 i 和主题 j 的相似度，w_i 为各指标对应的权重，$\sum_{i=1}^{3} w_i = 1$。

（1）文档重合度：根据李慧[27] 提出的方法确定文献子集范围，主题之间重合的文献数目越多，表明主题之间的相似度越高，计算不同主题之间文献子集的 Jaccard 系数（见公式（4））即为文档重合度。公式（4）中，$num(set_i \cap set_j)$ 表示主题 i 和主题 j 的文献子集取交集的数量，$num(set_i \cup set_j)$ 表示主题 i 和主题 j 的文献子集取并集的数量。

$$doc_{coincidence}(i,j) = \frac{num(set_i \cap set_j)}{num(set_i \cup set_j)} \qquad \text{公式(4)}$$

（2）主题词相似度：每个主题通常选择概率值较大的 N 个词语作为描述主题的特征词，以主题之间特征词的余弦相似度来度量主题词之间的相似度。如公式（5）所示，将所有主题下的特征词合并，n 为不重复

李慧，胡吉霞.一种基于图卷积自编码模型的多维度学科知识网络融合方法[J].图书情报工作,2020,64(18):114-125.

特征词总数，p_{im}表示特征词 m 在主题 i 词语分布中的权重，若主题 i 不包含特征词 m，则 p_{im} 值为 0。

$$feature_{sim}(i,j)=\frac{\sum_{m=1}^{n}(p_{im}*p_{jm})}{\sqrt{\sum_{m=1}^{n}p_{im}^{2}*\sum_{m=1}^{n}p_{jm}^{2}}}$$

公式(5)

(3)关键词相似度：将主题对应文献子集中文档包含的关键词转化为向量空间模型，计算关键词的余弦相似度，如公式(6)所示，tf_{ki}表示主题 i 的第 k 个关键词出现的频次。

$$keywords_{sim}(i,j)=\frac{\sum tf_{ki}\times tf_{kj}}{\sqrt{\sum(tf_{ki})^{2}}\times\sqrt{\sum(tf_{kj})^{2}}}$$

公式(6)

3.3.2 关键词子网络

统计所有文献中出现的关键词，若两个关键词在同一篇文献中出现，则这两个关键词存在一次共现关系，根据关键词同时出现的文献篇数即共现次数构建共词矩阵来挖掘关键词之间的相似性关系，共现次数越多表示关键词之间的相似度越高。

3.3.3 实体子网络

使用 LR 算法识别实体之间的相似关系，LR 是反映真实性的一种指标，定义为有约束条件下似然函数最大值和无条件约束下似然函数最大值之比，计算如公式(7)所示，这里一个实体出现的约束条件为另一个实体是否出现，即条件概率 $p(e_1|e_2)$，计算如公式(8)所示，表示实体 e_2 出现的情况下实体 e_1 出现的概率，若这两个实体经常一起出现即 LR 的值较大，则说明这两个实体之间具有较强的关联关系。基于此计算一个实体和其他实体的 LR 值，选取 LR 值较大的实体作为该实体的相似实体，以此生成实体相似度矩阵。

$$\mathrm{LR}(e_1,e_2)=\frac{p(e_1|e_2)}{p(e_1|not\ e_2)}$$ 公式(7)

$$p(e_1\mid e_2)=\frac{p(e_1,e_2)}{p(e_2)}$$ 公式(8)

根据相似度矩阵分别构建主题子网络、关键词子网络和实体子网络，作为学科知识网络 3 个维度的知识结构图，为后续网络融合工作奠定基础。

3.4 知识子网络融合

从不同维度构建的知识子网络中存在相同或者相似的知识节点，比如主题、关键词和实体网络节点可能表示同一种技术方法、名词术语或者研究对象。知识网络融合是将使用不同规则构建的知识子网络整合为一个更完整的知识网络，其关键是判断不同网络中两个知识单元是否描述同一对象，包括节点对齐和结构融合两个步骤。

3.4.1 节点对齐

本文参考论文[23]中提出的节点传递对齐方法，将任意大小的网络转为固定大小的网络结构，该算法分为 3 个步骤，框架如图 2 所示，伪代码见表 3。本文将3.3 节构建的 3 个知识子网络表示为 $G=\{G_1,G_2,G_3\}$，每个知识子网络结构可表示为 $G_P=(V_P,E_P,A_P,X_P)$，V_P 表示节点的集合，E_P 表示连边的集合，A_P 表示子网络 G_P 的邻接矩阵，X_P 表示节点的属性特征矩阵。

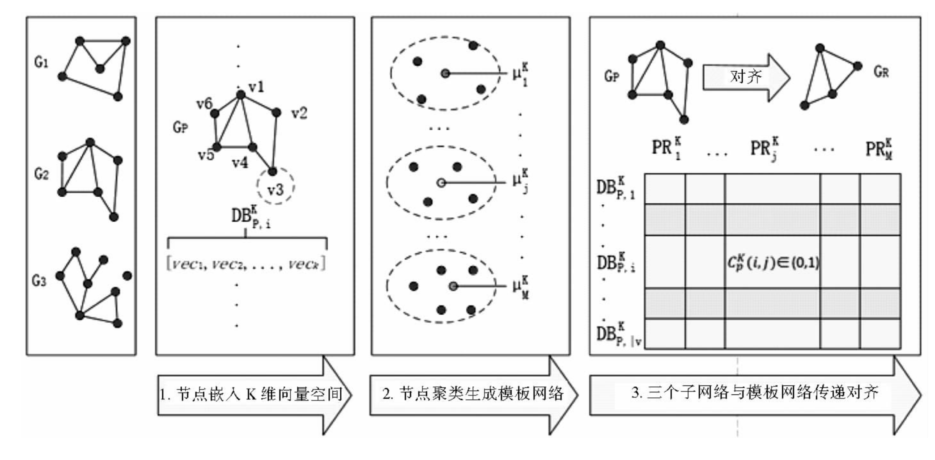


图 2 节点传递对齐算法框架

(1)节点嵌入 K 维向量空间。假设任意一个知识子网络 G_P 包含 n 个节点，使用节点嵌入方法将知识子网络中的每个知识单元映射到一个 K 维向量空间进行向量化表示。第 P 个子网络中第 i 个节点的 K 维特征向量记作 $DB_{p,i}^{K}=\{vec_1,vec_2,\cdots,vec_K\}$，所有知识单元的向量可用集合 $R^K=\{R_1^K,R_2^K,\cdots,R_N^K\}$ 表示，N 为知识单元总数。

(2)节点聚类生成模板网络。使用 Kmeans 聚类

算法将所有子网络的节点聚成 M 个类，通过最小化目标函数（见公式(9)）得到 M 个聚类中心 $PR^K=\{\mu_1^K,\mu_2^K,\cdots,\mu_M^K\}$，每个聚类中心节点用 K 维向量进行表示，M 个聚类中心构成一个模板网络。

$$\arg\min_{\Omega}\sum_{j=1}^{M}\sum_{R_i^K\in c_j}\| R_i^K-\mu_j^K\|^2 \quad \text{公式(9)}$$

（3）所有子网络结构与模板网络传递对齐。分别计算每个子网络中的每个节点与模板网络节点集合的距离矩阵，将第 P 个知识子网络与模板网络的 K 维节点向量的距离矩阵表示为 D_P^K，见表 1，子网络的第 i 个节点与模板网络的第 j 个节点的欧式距离可由公式(10)计算得出。

$$D_P^K(i,j)=\sqrt{\sum_K \| DB_{p,i}^K-\mu_j^K\|^2} \quad \text{公式(10)}$$

表 1　子网络与模板网络的距离矩阵

	μ_1^K	μ_2^K	…	μ_j^K	…	μ_M^K
R_1^K	$D_P^K(1,1)$	$D_P^K(1,2)$	…	$D_P^K(1,j)$	…	$D_P^K(1,M)$
R_2^K	$D_P^K(2,1)$	$D_P^K(2,2)$	…	$D_P^K(2,j)$	…	$D_P^K(2,M)$
…	…	…	…	…	…	…
R_i^K	$D_P^K(i,1)$	$D_P^K(i,2)$	…	$D_P^K(i,j)$	…	$D_P^K(i,M)$
	…	…	…	…	…	…
R_n^K	$D_P^K(n,1)$	$D_P^K(n,2)$	…	$D_P^K(n,j)$	…	$D_P^K(n,M)$

对齐矩阵 C_p^K 表示为一个二值矩阵，可以从距离矩阵 D_P^K 中推导出来：在距离矩阵 D_P^K 中，若第 i 行第 j 列为第 i 行的最小值，则对齐矩阵对应位置的值为 1，否则为 0，如公式(11)所示：

$$C_p^K(i,j)=\begin{cases}1, & \text{若 } D_P^K(i,j)\text{ 为第 i 行最小值}\\ 0, & \text{其他}\end{cases} \quad \text{公式(11)}$$

在对齐矩阵中，每行仅有一个元素为 1，其余为 0，表示子网络中的每个节点都只对应模板网络中的一个节点，与模板网络中特定节点对应的子网络节点可能有多个，因为模板网络节点是聚类产生的，子网络节点之间具有相似性，相同或相似节点与同一模板节点对齐是合理的，如表 2 所示。另外，如果两个子网络中的节点都与模板网络中相同的节点对齐时，这两个子网络的节点也是对齐的，因此这种对齐关系是传递的。

表 2　子网络与模板网络的对齐矩阵

	μ_1^K	μ_2^K	…	μ_j^K	…	μ_M^K
R_1^K	0	1	…	0	…	0
R_2^K	1	0	…	0	…	0
…	…	…	…	…	…	…
R_i^K	0	0	…	1	…	0
…	…	…	…	…	…	…
R_n^K	0	0	…	1	…	0

得到对齐矩阵 C_p^K 后，第 P 个子网络添加自环的邻接矩阵为 $\tilde{A}_P$，即 $\tilde{A}_P=A+I$，I 为单位矩阵，节点属性特征矩阵为 X_p，利用公式(12)和公式(13)可计算对齐后各子网络的邻接矩阵 $\bar{A}_P^K$ 和特征矩阵 $\bar{X}_P^K$。

$$\bar{A}_P^K=(C_P^K)^T(\tilde{A}_P)(C_P^K) \quad \text{公式(12)}$$

$$\bar{X}_P^K=(C_P^K)^T X_p \quad \text{公式(13)}$$

节点传递对齐算法的伪代码如表 3 所示：

表 3　节点传递对齐算法伪代码

Input：3 个子网络 $G=\{G_1,G_2,G_3\}$ 及网络结构 $G_P=(V_P,E_P,A_P,X_P)$

Output：各网络对齐后的邻接矩阵 $\bar{A}$ 及特征矩阵 $\bar{X}$

```
method node_aligned                                  /* 节点传递对齐算法 */
  for P in [1,3] do                                  /* 节点嵌入 */
     for V_i in V_P do
        DB^K_{p,i} ← {vec_1, vec_2, …, vec_K}         /* 节点 i 的 K 维向量表示 */
     end for
  end for
  R^K ← {R^K_1, R^K_2, …, R^K_n}                      /* 所有节点的向量表示 */
  (C_1, C_2, …C_M) ← Kmeans                           /* 聚类 */
  for C_i in (C_1, C_2, …C_M) do                      /* 计算聚类中心 */
     μ^K_i ← (1/num) Σ_j^num DB^K_{p,i}               /* 聚类中心：类中节点向量的均值 */
  end for
  for i in [1,M] do                                  /* 节点对齐 */
     for j in [1,n] do
        D^K_P(i,j) ← sqrt(Σ_K ‖DB^K_{p,i} − μ^K_j‖^2)   /* 距离矩阵 */
     end for
  end for
  for i in [1,M] do                                  /* M 个类 */
     for j in [1,n] do                               /* n 个节点 */
        if D^K_P(i,j) == min D^K_P(i,) do /* 若 D_P^K (i,j)是第 i 行最小值 */
           C^K_P(i,j) ← 1
        else do
           C^K_P(i,j) ← 0
        end if
     end for
  end for
  for P in [1,3] do                                  /* 对齐后的邻接矩阵和特征矩阵 */
     Ā^K_P ← (C^K_P)^T (Ã_P) (C^K_P)
     X̄^K_P ← (C^K_P)^T X_p
  end for
  return Ā^K_P, X̄^K_P
end method
```

3.4.2　结构融合

将网络节点结构信息的表示学习问题转化为词语的表示学习问题，本文利用神经网络语言模型挖掘网络节点属性的深层语义信息，再利用属性相似性对网络结构进行刻画。经过传递对齐后的所有子网络均有 M 个节点，N 个子网络共有 N * M 个节点。将所有节

点按序排列后，运用基于图卷积运算的自编码神经网络模型进行联合训练，得到综合节点属性信息和结构信息的节点表示向量。该算法包括3个步骤，框架流程如图3所示，具体描述如下：

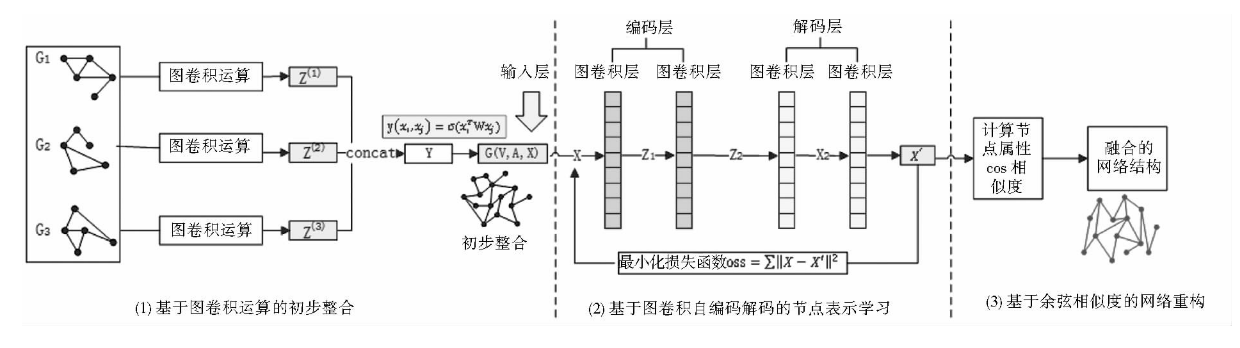


图3　子网络融合算法流程

（1）基于图卷积运算的初步整合。对任意一个子网络 $G_P = (V_P, E_P, \bar{A}_P, \bar{X}_P)$，令 $\tilde{A}_P = \bar{A}_P + I_M$，M为节点个数，$I_M$ 为单位矩阵，对角矩阵 D_P 为邻接矩阵的度矩阵，对角元素 $D_{ii}^P = \sum_j \tilde{A}_{P(i,j)}$，表示与节点i相连的边的数量。对网络 G_P，通过一个两层的图卷积运算得到节点嵌入 $Z^{(P)}$，如公式(14)所示。

$$Z^{(P)} = f(\bar{X}_P, \bar{A}_P, W_0^P, W_1^P) = Softmax(\hat{A}_P\ Relu(\hat{A}_P\ \bar{X}_P W_0^P) W_1^P) \quad 公式(14)$$

其中，$\hat{A}_P = D_P^{-1/2}\ \tilde{A}_P\ D_P^{-1/2}$ 表示对邻接矩阵 $\tilde{A}_P$ 进行归一化，W_0^P，W_1^P 表示第一层和第二层的权重矩阵，*Relu* 为非线性激活函数。经过计算，每个子网络都会得到一个节点嵌入 $Z^{(P)}$，将所有子网络的节点嵌入按顺序排列起来，生成一个包含N＊M个节点的嵌入矩阵Z，计算sigmoid函数值 $y(x_i, x_j) = \sigma(z_i^T W_{z_j})$，*W* 为超参数权重矩阵，$z_i$，$z_j$ 分别表示嵌入矩阵Z中节点 i 和节点 j 的嵌入向量，选择合适的阈值，大于阈值的节点之间建立连边，得到初步整合的网络结构。

（2）基于图卷积自编码的节点表示学习。图卷积神经网络的卷积操作可以聚合节点的属性和连接信息，自编码网络可以在无监督情况下进行表示学习，本文将图卷积神经网络和自动编码器相结合，构建图卷积自编码网络模型。根据输入网络的节点和连接信息设置卷积参数，通过最小化损失函数来指导网络训练。如图3中第(2)部分所示，该网络分为编码器和解码器两部分，编码器为聚合网络节点属性和连接信息的卷积编码层，解码器为对卷积特征进行重建的解码层，输出是在设法重建节点的输入属性，损失函数即为重建损失，如公式(15)所示，当输出与输入的差异越小时，表明网络的学习能力越强。

$$loss = \frac{1}{N*M} \sum_{i=1}^{N*M} \| X_i - X_i' \|^2 \quad （公式15）$$

编码器和解码器的结构都包含了两个卷积层，在本文的实证研究部分，输入节点数为213个，节点特征维数为100，其权重参数设置如表4所示。通过不断调整参数来最小化损失函数，以期从网络结构中获取更多的信息量，并将最后一次的输出信息 X' 作为节点的属性特征向量，其与输入特征有着相同的维数。

表4　图卷积自编码网络模型参数设置

	卷积层	输入	输出	激活层
编码	1	213	64	Softmax
	2	32	100	Sigmoid
解码	1	213	32	Sigmoid
	2	64	100	Softmax

（3）基于余弦相似度的网络重构。将图卷积自编码网络模型的输出信息作为网络节点的向量表示，利用向量空间的余弦相似度（见公式(16)，*K* 为向量的维数）计算节点之间的相似程度，并确定合适的阈值，以此来构建最终的学科知识网络。

$$cos_sim = \frac{\sum_{i=1}^{K}(x_i * y_i)}{\sqrt{\sum_{i=1}^{K} x_i^2} * \sqrt{\sum_{j=1}^{K} x_j^2}} \quad 公式(16)$$

4　实证研究

“人工智能”作为目前计算机领域最热门的研究方向之一，已经成为推动社会发展不可或缺的技术资源，其跨学科性和广泛的应用前景使得相关研究成果层出不穷，是近几年国内外科学研究的热点问题。为了帮助科学研究者充分了解该领域学科知识结构，同时验证本文提出的知识网络融合方法的有效性，本文选择人工智能领域的中文文献作为实验数据来源，挖掘学科知识结构并进行可视化，最后对实验结果展开分析。

4.1　数据获取与预处理

本文检索人工智能领域近十年的中文文献，删除

少量关键词和摘要信息不完整的数据后，共得到文献 10 224 篇，如表 5 所示。利用 Hanlp 软件包[28]对文献的标题和摘要进行分词、提取双连词、去停用词、词性标注等预处理，生成建模需要的语料库。

表 5 数据来源

数据类型	检索时间范围	数据库来源	检索表达式	来源类别	文献类型	检索结果
中文文献	2009/01/01——2019/01/01	中国知网 CNKI	(SU = 人工智能 or TI = 人工智能) and (KY = 人工智能 or KY = AI)	SCI、EI、核心期刊、CSSCI 和 CSCD	期刊	10 224 篇

4.2 实验设置与结果分析

4.2.1 实验参数设置

(1)LDA 主题抽取。LDA 的相关算法已经很成熟，通过计算链接困惑度确定最优主题数 T 为 53，其他参数根据参考文献[29]和经验值设定，具体如表 6 所示：

表 6 LDA 模型参数说明

模型参数	参数说明
α	文本集在潜在主题上的狄利克雷先验，α = 50/T
β	潜在主题在特征词集上的狄利克雷先验，β = 0.02
T	最优主题数 53
niters	Gibbs 抽样迭代次数，niters = 1000
twords	主题下特征词个数，twords = 30

(2)知识子网络构建。利用前文提到的知识单元抽取和相似度计算方法，计算各维度知识单元的相似度并以此构建知识子网络，经过多次实验对比，各维度知识单元数目及关联阈值等参数最终设定如表 7 所示。其中，关键词提取词频 20 以上、累计占比 25%。

表 7 知识子网络相关参数

知识维度	知识单元数目	关联阈值	关联关系(对)
主题	53	0.311	396
关键词	152	20(词频)	2 692
实体	100	0.221	701

(3)节点对齐。本文用词向量来表示知识单元的属性特征，在主题抽取部分，对各主题含义进行了人工概括总结，可能出现部分表示主题的词语并未在语料库中出现，Word2Vec 模型无法学习此类词语的向量表示，同时考虑到程序实现的简单性原则，因此本文选择 gensim 软件包的 FastText[30]模型学习知识单元的词向量表示。高纬度的词向量可以更丰富地表示词组的语义信息，同时也会增加神经网络模型参数的数量而导致过拟合，根据参考文献[31-34]将词向量维数设为 100。

根据词向量对知识单元进行 K-means 聚类，通过轮廓系数法确定最优聚类数目 M，如图 4 所示。最初抽取的知识单元总数为 305，考虑到最终网络节点个数不应该超过知识单元总数，且为了降低信息损失，最终节点个数不能太小，因此将最小聚类数目设为 50，M 大于 90 的不再考虑，最终聚类数目 M 设为 71。3 个子网络节点对齐后，最终会得到 213 个节点的 100 维向量表示，再计算对齐后的邻接矩阵和特征矩阵，作为图卷积自编码网络的输入，进行节点属性特征表示学习。

4.2.2 实验结果分析

(1)单个子网络分析。根据前文所述子网络构建方法，以 3 个维度抽取的知识单元构建知识关联子网络，利用 Pajek 软件进行可视化，根据节点的度和连接权重对节点分类，以不同颜色进行区分，如图 5 - 图 7 所示。其中，度相同的节点有相同的颜色，节点越大表示节点越重要，对应的研究内容通常用来表示一个领域的研究热点。

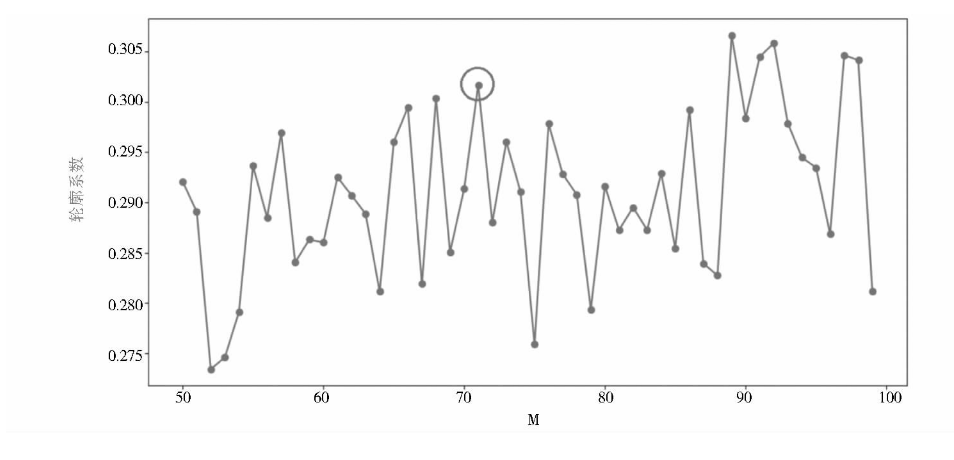


图 4 轮廓系数法确定聚类数目 M

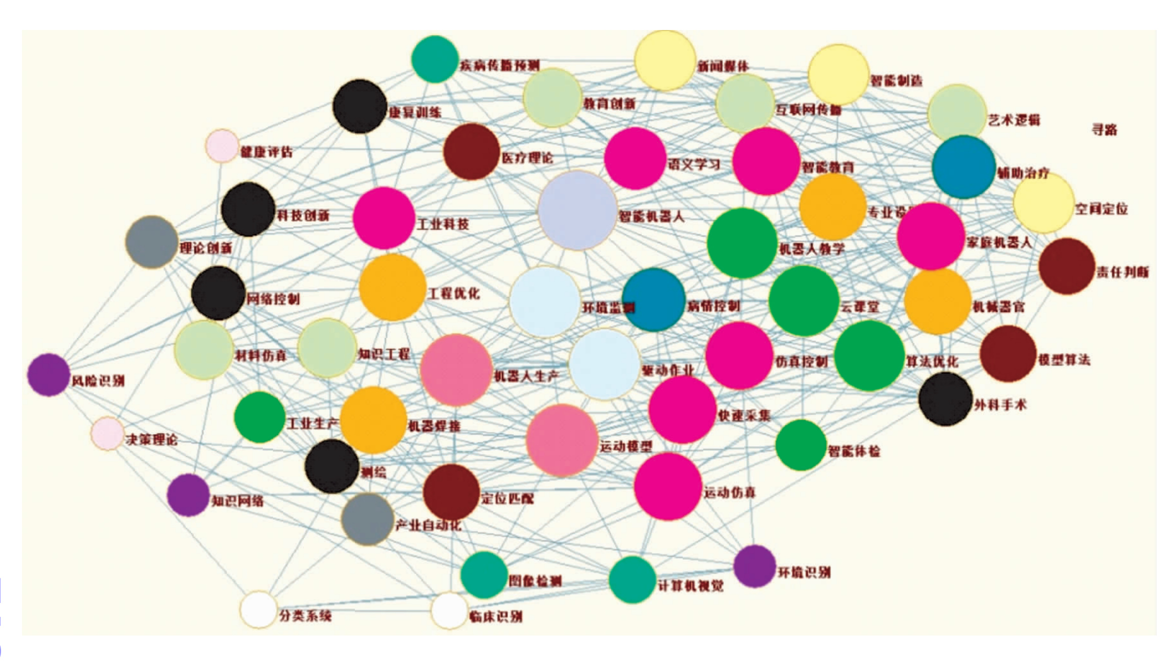


图 5　主题子网络结构

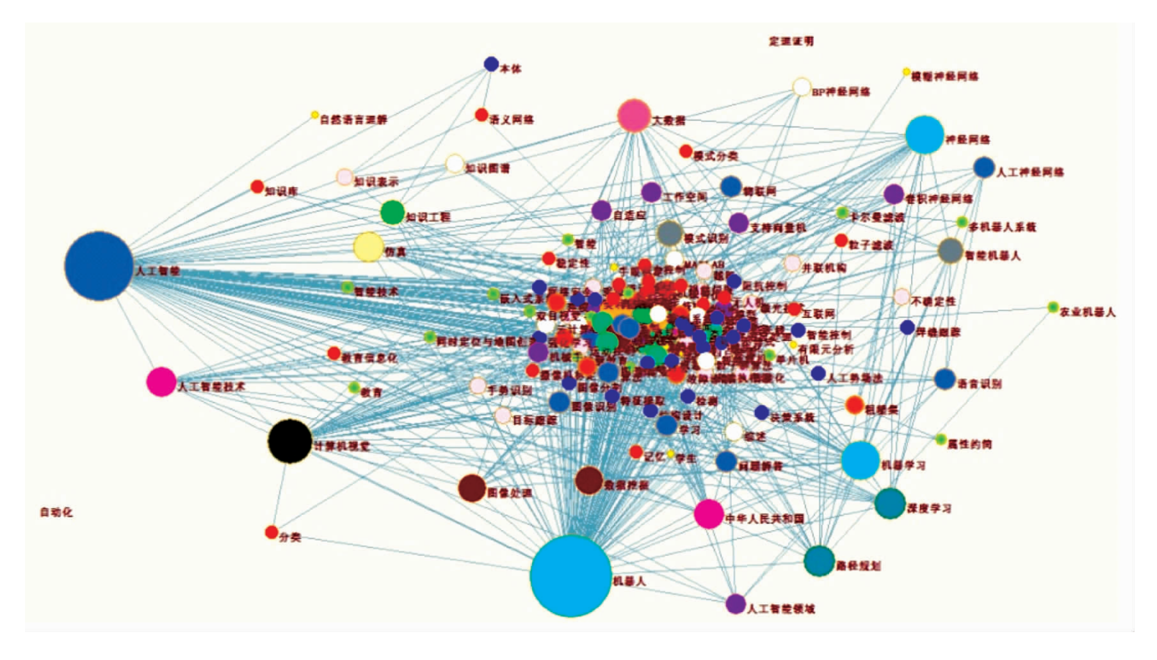


图 6　关键词共现子网络结构

从图 5 可以看出，主题网络中节点大小比较接近，主题倾向于从宏观层面表示领域研究内容，如智能机器人、环境监测、驱动作业、运动模型和机器人生产。从图 5 还观察到，人工智能领域的主题主要侧重于技术应用层面的研究，重点关注智能制造、智能体检、家庭机器人、云课堂和决策理论，涉及制造业、医疗、智能家居和教育等传统行业的应用，部分涉及语义学习、空间定位、模型算法和计算机视觉等技术的研究，说明目前人工智能领域的技术已逐渐趋于成熟。另外，主题网络中存在一个孤立节点“寻路”，该主题理应与空间定位、定位匹配、运动模型和运动仿真等主题直接或间接关联，表示与机器人运动路径相关的研究。图 6 为关键词共现网络，与主题网络相比，更侧重于从细粒度描述领域研究内容，如支持向量机、卡尔曼滤波、粗糙集等具体算法。关键词既有自然科学领域对人工智能技术本身的研究，又有社会科学领域对人工智能的应用场景，如大数据、神经网络、机器学习、深度学习、路径规划、图像处理和计算机视觉等炙手可热的关键技术，以及物联网、智能机器人、农业机器人、决策系统和知识工程等相关领域的应用。但最关键的研究内容依

然是智能机器人（"人工智能"和"机器人"两个关键词所占比重最大）的研究，这与最热主题为"智能机器人"相一致。从图 7 实体关联网络可以观察到，实体更多关注领域内表示研究对象、行业、团体、工具等实物性内容，如地图、图像、机器、课程、教师、文化、制造业、新媒体、会议、专业委员会和专家系统等，为了能与主题和关键词网络更好地融合，在筛选实体时也包含一些关键技术，如深度学习、智能控制、模式识别等。图 5 - 图 7 中都包含极少数孤立节点，如主题寻路、关键词定理证明和自动化、实体机器人运动学等。在主题网络中，即使综合前文所述的 3 种方法对主题的相似度进行计算，依然存在部分关联关系未被挖掘出来，因此，对 3 个网络进行融合及重构以期挖掘出完整的学科知识结构显得十分必要。

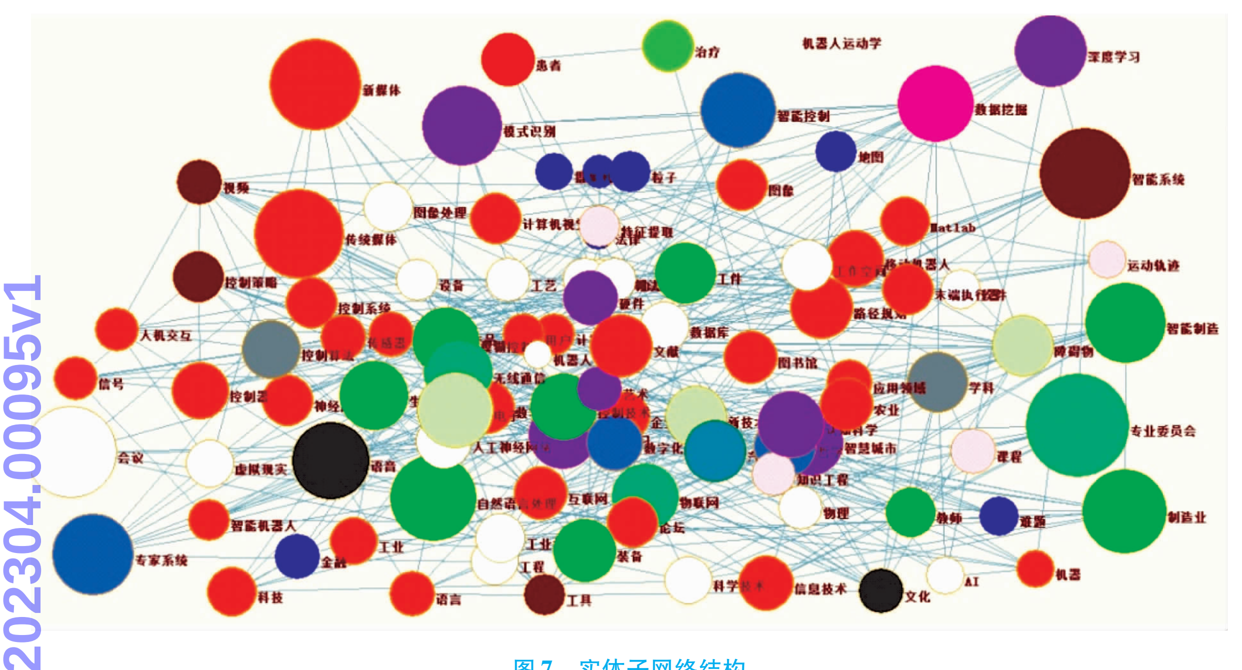


图 7　实体子网络结构

（2）整合网络分析。利用 3.4 节提出的图卷积自编码网络模型学习网络节点的向量表示，神经网络模型的学习率为 0.01，迭代次数为 1 000，经过多次实验，平均准确率在 0.86 左右。利用训练出的节点属性特征重新计算节点的相似程度，通过筛选共保留 3 721 对关系，并根据聚类结果和对齐矩阵对同一类含义接近的节点进行合并，对节点名称进行概括总结，构建的学科知识网络如图 8 所示。从图 8 中可以看出，融合后的知识网络结构中不存在孤立节点，网络中包含宏观的主题节点如风险识别、仿真技术、机器人、图像处理、数据挖掘等，也包含微观层次的关键词如卷积神经网络、人工神经网络、三元组、跟踪算法等，另外，表示实体的图片、视频、患者、车辆、地图等也囊括其中。图中包含 3 种类型的节点，都可看作人工智能领域的知识单元，与传统单个子网络相比，融合的知识网络结构能够更全面地反映学科的研究内容和知识结构，具体做如下分析：

第一，研究热点。观察到图中较大的节点有抓取姿态、机械臂、机器人教学、机器人、农业机器人、工业机器人、跟踪算法、路线、神经网络、人工神经网络、卷积神经网络、图片、地图、图像检测、图像处理、最优化、仿真技术、信息资源、风险识别、环境识别等，涉及人工智能领域智能机器人、路径规划、深度学习、计算机视觉、信息检索以及人工智能用于指导决策等相关问题的研究，表明这些内容是领域关注的热点，经参考《2019 人工智能发展报告》[35] 相关介绍，笔者发现与人工智能领域的重点研究内容基本一致。

第二，知识关联。整合后的知识网络中，热点研究问题和关键技术与其他知识节点联系紧密，如与机器人、图像检测、图片、神经网络等节点相连的知识节点较多。另外，在网络边缘部分可以看到，涉及智能教育、机器学习、智能机械、医疗机器人和仿真等知识单元之间的连接关系比较紧密，与其他知识簇的联系比较稀疏，在网络中形成一个个小的社团。

（3）讨论。本文提出的知识网络构建模型是一种无监督学习方法，人工智能领域没有权威的知识网络可供参考，传统的评价指标如准确率、召回率、精度等也不适合用来验证本文方法的有效性。查阅相关文献[36]，从以下 3 个方面分析所提方法的有效性：

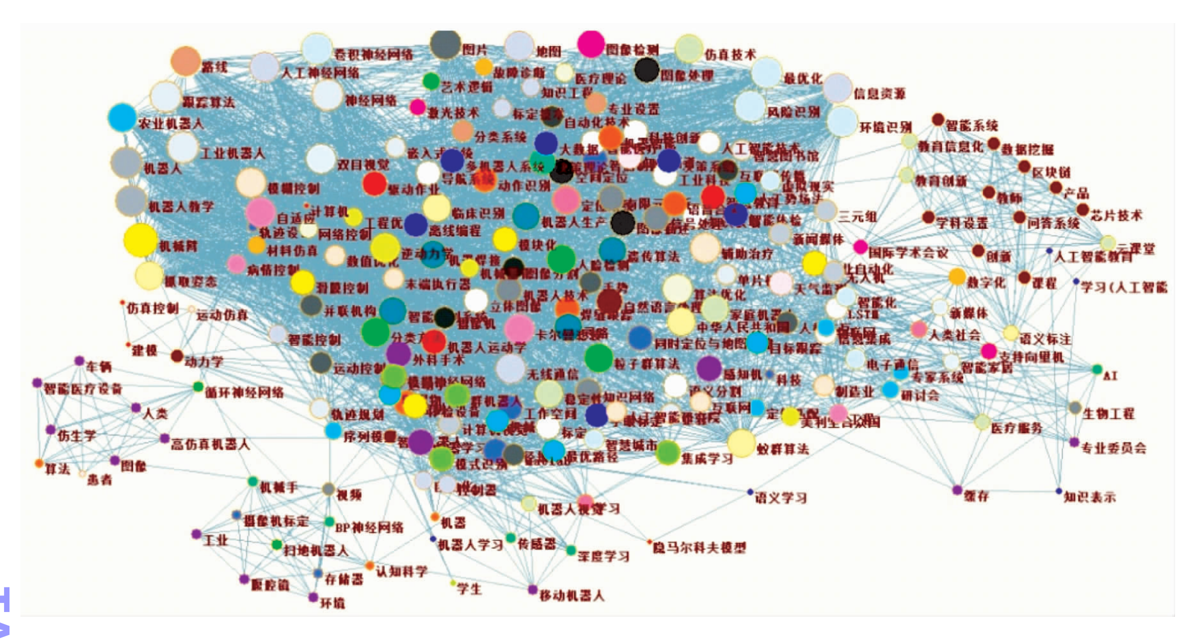


图8　融合的网络结构

第一，子网络节点重复率。对比3个子网络图中的节点可以发现，主题、关键词和实体3种类型的节点存在一定的重复性，用重复节点数与所有节点数的比值来表示节点重复率，如公式(17)所示：

$$节点重复率 = \frac{重复节点数}{节点总数} \qquad 公式(17)$$

统计知识单元集合中的重复词和同义词(如“人工智能”和“AI”)，本文构建的3个子网络共305个节点，其中两个网络相重复的节点数48个，两节点重复率为15.74%，3个网络中相重复的节点数有4个，分别是智能制造、计算机视觉、智能机器人和知识工程，从含义来看，这几个知识单元既能表示研究主题，也可作为文献关键词，还可用来表示相关实体概念，其重复率为1.312%。可见，两个网络的节点重复率较高，如果直接通过节点词向量计算节点相似度的方法进行整合，会存在部分节点重复，从而给知识网络结构带来一定的冗余性。经过节点聚类后，相同或相似的知识单元被聚为同一类，对同一类别的知识单元进行再总结，可有效降低知识单元的重复率。

第二，融合网络新增节点占比。结合领域背景和知识子网络知识单元的内容，对聚类后的知识单元名称进行总结，保留多数原知识节点，对相似节点含义重新概括，包含213个节点，其中，原节点150个，节点保留率70.42%，新增节点63个，占比29.58%，如公式(18)所示。通过聚类方法对齐节点，既保留了大部分原始单一网络中的知识单元，又通过新增节点对原知识单元进行了概括补充，能够更充分地展示一个领域的研究内容。

$$节点保留(新增)率 = \frac{融合网络中原节点(非原节点)数}{融合网络节点总数}$$

公式(18)

第三，融合网络新增连边占比。原知识子网络中的关系只包含同类型知识单元之间的联系，比如主题和主题之间，融合后的知识网络包含知识子网络内部的结构和子网络之间的关联关系。融合的知识网络共3 721条边，统计新增节点之间的连边、原节点之间的新增连边以及原节点和新增节点之间的连边共2 324条，占比62.46%，即有37.54%的关系来自原知识网络。因此，文中提出的图卷积自编码模型很好地聚合了原始网络结构，同时挖掘出更多的知识关联关系。

通过上述分析，融合的知识网络在领域研究热点发现、知识关联关系挖掘等方面表现良好，利用节点聚类和图卷积网络自编码模型能够聚合子网络结构和节点的属性信息，能较全面准确地揭示领域知识单元之间的关联关系。同时，本文构建神经网络无监督模型学习知识节点的特征信息，不需要标记数据，对其他应用场景同样适用，对异构网络融合具有借鉴价值。

5　结论

本文提出了一种融合主题、关键词和实体的学科知识网络构建方法。首先利用自然语言处理方法对中文语料进行预处理，采用LDA、TF-IDF方法抽取人工智能领域的主题和实体，提取语料中的关键词，其次，基于语义相似度和关键词共现分析构建学科知识子网

络,然后设计了基于图卷积的自编码网络模型学习知识单元向量表示方法,最后利用余弦相似度重构整个学科知识网络,达到挖掘学科知识结构的目的。通过对人工智能领域知识网络的分析和讨论,证明方法的有效性和准确性。

对学科领域知识点进行抽取并有效组织,可以帮助科学研究者快速了解领域研究热点和知识结构。现有知识网络构建方法涉及多维度知识融合的方法较少,本文提出的知识网络融合方法,不但能捕捉到知识单元的语义信息,还对子网络中节点的结构信息也进行了聚合,这种无监督的知识单元表示学习方法效率更高,其学习到的节点向量具有通用性,可用于解决知识单元聚类、分类等问题。

本文的不足之处在于:对主题和聚类后的知识单元含义进行了人工总结,存在一定的主观性;通过计算余弦相似度进行网络重构,增加了工作量,未来将尝试设计更先进的算法对节点的连边进行预测;实证研究部分只对中文的文献数据进行了验证,抽取的学科知识内容可能不全面,未来考虑融合多种数据源构建学科知识网络。

参考文献:

[1] 赵蓉英. 论知识网络的结构[J]. 图书情报工作,2007,51(9):6-10.
[2] 顾东蕾. 论学科知识网络的理论基础[J]. 图书情报工作,2008,52(9):32-35,73.
[3] 王晓光. 科学知识网络的形成与演化(Ⅰ):共词网络方法的提出[J]. 情报学报,2009,28(4):599-605.
[4] SEUFERT A, KROGH G, BACH A. Towards knowledge networking[J]. Journal of knowledge management, 1999, 3(3):180-190.
[5] 赵蓉英. 知识网络及其应用[M]. 北京:北京图书馆出版社,2007:8-58.
[6] 顾东蕾. 论学科知识网络[J]. 情报杂志,2008(9):50-55.
[7] 寇继虹. 学科领域知识网络的可视化构建研究——以竞争情报为例[J]. 信息资源管理学报,2015,5(3):71-77.
[8] 肖冬平. 知识网络研究综述[J]. 重庆工商大学学报(自然科学版),2006(6): 617-623.
[9] 王曰芬,李冬琼,余厚强. 生命周期阶段中的科学合作网络演化及高影响力学者成长特征研究[J]. 情报学报,2018,37(2):121-131.
[10] 潘有能,谭健. 普赖斯奖得主的科学合作网络研究[J]. 图书情报工作,2012, 56(16):80-84.
[11] 邱均平,周毅. 基于作者共被引的馆藏资源深度聚合模式与服务探析——以CSSCI中图书情报领域本体研究为例[J]. 图书情报工作,2014,58(7):19-24.
[12] 侯剑华. 国际科学计量学研究前沿的可视化探测——基于《Scientometrics》期刊文献共被引网络的分析[J]. 现代情报,2012,32(10): 61-65.
[13] 姜春林,张帆,唐悦. 我国部分科学学期刊共被引网络特征研究[J]. 情报杂志,2010,29(4):10-15,25.
[14] 刘秋霞,吴汉卿,黄正来. 基于全球文献计量的小麦响应气候变暖的研究[J]. 中国农学通报,2019,35(23):142-151.
[15] 罗润东,滕宽,李超. 2018年中国经济学研究热点分析[J]. 经济学动态,2019 (4):80-98.
[16] 张怡青,王高玲. 基于知识图谱的国内外健康管理研究对比分析[J]. 中国全科医学, 2019,22(9):1112-1118.
[17] 吕鹏辉,张士靖. 学科知识网络研究(Ⅰ)引文网络的结构、特征与演化[J]. 情报学报, 2014,33(4):340-348.
[18] 吕鹏辉,张凌. 学科知识网络研究(Ⅱ)共被引网络的结构、特征与演化[J]. 情报学报, 2014,33(4):349-357.
[19] 赵一鸣,吕鹏辉. 学科知识网络研究(Ⅲ)共词网络的结构、特征与演化[J]. 情报学报, 2014,33(4):358-366.
[20] 关鹏,王曰芬,曹嘉君. 整合主题的学科知识网络构建与演化分析框架研究[J]. 情报科学,2018,36(9):3-8.
[21] 王曰芬,王金树,关鹏. 主题-主题关联的学科知识网络构建与演化分析[J]. 情报科学, 2018,36(9):9-15,102.
[22] 何劲,关鹏,王曰芬. 作者-主题关联的学科知识网络构建与演化分析[J]. 情报科学, 2019,37(1):56-62,67.
[23] BAI L,JIAO Y,CUI L,et al. Learning aligned-spatial graph convolutional networks for graph classification[C]//ECML PKDD 2019. Machine learning and knowledge discovery in databases. Würzburg: Springer,2019:464-482.
[24] 胡玉宁,胡观伟. 多源主题融合的科学知识结构模型构建与实证研究[J]. 情报理论与实践, 2019,42(7):100-105.
[25] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research,2003(3) :993-1022.
[26] MICHAEL K. The lokahi prototype: toward the automatic extraction of entity relationship models from text[C]// Proceedings of the AAAI 2019 spring symposium on combining machine learning with knowledge engineering (AAAI-MAKE 2019). Palo Alto:Stanford University, 2019:121-126.
[27] 李慧,田亚丹. 一种层次化的科学知识结构发现方法[J]. 图书情报工作, 2018,62(13):92-102.
[28] 上海林原信息科技有限公司. HanLp[EB/OL]. [2020-01-24]. http://www.hanlp.linrunsoft.com/.
[29] 王鹏,高铖,陈晓美. 基于LDA模型的文本聚类研究[J]. 情报科学,2015 (1):63-68.
[30] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[C]//The conference on transactions of the Association for Computational Linguistics. Prangue: ACL, 2017:135-146.
[31] 冶忠林,赵海兴,张科,等. 基于多源信息融合的分布式词表示学习[J]. 中文信息学报,2019,33(10):18-30.
[32] 冶忠林,赵海兴,张科,等. 基于描述约束的词表示学习[J]. 中文信息学报, 2019,33(4):29-36.

[33] 赖文辉,乔宇鹏.基于词向量和卷积神经网络的垃圾短信识别方法[J].计算机应用,2018,38(9):2469-2476.

[34] WU C, GAO R, ZHANG Y, et al. PTPD: predicting therapeutic peptides by deep learning and word2vec[J]. BMC bioinformatics, 2019,20(15):87-108.

[35] 清华大学-中国工程院知识智能联合研究中心,北京国人工智能学会吴文俊人工智能科学技术奖评选基地.2019 人工智能发展报告[EB/OL].[2020-01-24]. https://www.sohu.com/a/360140139_468661.

[36] LU R, FEI C, WANG C, et al. HAPE: A programmable big knowledge graph platform[J]. Information sciences, 2020(509): 87-103.

作者贡献说明:
李慧:提出研究思路和论文修改意见;
胡吉霞:负责实验实施,论文撰写和修改。

Multi-Dimensional Subject Knowledge Network Fusion Method Based on Graph Convolution Self-Encoding Model

Li Hui Hu Jixia
School of Economic & Management, Xidian University, Xi'an 710126

Abstract: [**Purpose/significance**] Aiming at the problem that the knowledge network containing a single type of knowledge unit cannot fully reflect the knowledge structure of the subject, a method of integrating knowledge network structure in different dimensions is proposed to provide a reference for the knowledge structure mining in the subject area. [**Method/process**] This paper used LDA and TF-IDF methods to extract subject knowledge units, and then used semantic similarity and keywords co-occurrence analysis methods to construct three subject knowledge sub-networks: topics network, keywords network and entities network, and adopted spatial nodes transfer alignment align the nodes of the sub-networks, then designed a self-encoding model based on the graph convolution operation to represent the knowledge nodes, and finally reconstructed the disciplinary knowledge network by calculating the cosine similarity. [**Result/conclusion**] The experimental part takes the field of artificial intelligence as an example to construct a subject knowledge network that integrates topics, keywords, and entities and conducts analysis. The experimental results show that the method proposed in this article can effectively reveal the research content and knowledge structure of the subject area, and provide a useful reference for the discovery and organizational research of subject knowledge.

Keywords: network fusion knowledge structure node alignment graph convolutional neural network self-coding model